REFERENCE RESOLUTION USING SEMANTIC PATTERNS IN JAPANESE NEWSPAPER ARTICLES

Takahiro Wakao

University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello St, Sheffield S1 4DP, UK
Email: *t.wakao@dcs.shef.ac.uk*

1 INTRODUCTION

Reference resolution is one of the important tasks in natural language processing. In Japanese newspaper articles, pronouns are not often used as referential expressions for company names, but shortened company names and *dousha* ("the same company") are used more often (Muraki *et al.* 1993). Although there have been studies of reference resolution for various noun phrases in Japanese (Shibata *et al.* 1990; Kitani 1994), except Kitani's work, they do not clearly show how to find the referents in computationally plausible ways for a large amount of data, such as a newspaper database. In this paper[1], we determine the referents of *dousha* and their locations by hand, and then propose one simple and two heuristic methods which use semantic information in text such as company names and their patterns, so as to test these three methods on how accurately they find the correct referents.

Dousha is found with several particles such as "*ha*", "*ga*", "*no*", and "*to*" in newspaper articles. Those which co-occur with *ha* and *ga* are chosen for the data since they are the two most frequent particles when *dousha* is in the subject position in a sentence. Typically, *ha* marks the topic of the sentence and *ga* marks the subject of the sentence. A typical use of *dousha* is as follows:

> Nihon Kentakii Furaido Chikin ha,
> Japan Kentucky Fried Chicken ha,
>
> sekai saidai no piza chien,
> world's largest pizza chain store,
>
> Piza Hatto to teikei wo musubi,
> Pizza Hut to tie-up establish,
>
> kotoshi gogatsu kara zenkoku de
> starting May this year, nation-wide,
>
> takuhai piza chien no tenkai wo
> pizza delivery chain store extension
>
> hajimesu to happyou shita.
> begin announced.
>
> sarani *dousha ha* furaido chikin no
> Moreover, the same company fried chicken
> of
>
> takuhai saabisu nimo noridasu.
> delivery service as well will start.

A rough translation is:
"Kentucky Fried Chicken Japan announced that it had established a tie-up with the world largest pizza chain store, Pizza Hut, and began to expand pizza delivery chain stores nation-wide starting in May this year. Moreover, *the company* will start delivery of fried chicken as well."

Dousha in the second sentence refers to Kentucky Fried Chicken Japan as "*the company*" does in the English translation. As shown in this example, some articles contain more than one possible referent or company, and the reference resolution of *dousha* should identify the referent correctly.

2 LOCATIONS AND CONTEXTS OF THE REFERENTS

Most of the Japanese newspaper articles examined in this study are in the domain of Joint-Ventures. The sources of the newspaper articles are mostly *the Nikkei* and *the Ashahi*. The total number of the articles is 1375, and there are 42 cases of *dousha* with *ga* and 66 cases of *dousha* with *ha* in the entire set of articles.

The following tables, **Table** 1 and **Table** 2, show the locations and contexts where the referents of both subsets of *dousha* appear.

[1]This paper was written when the author was at the Computing Research Laboratory of New Mexico State University. The author has been at University of Sheffield since January 1994.

Table 1 Locations and contexts of the referents of *dousha* with *ga*

dousha with *ga*		
location	context	number of cases
Within the same sentence		19
Subject	company name + *ha*	7
	part of the subject ∗	1
Non-subject	company name + *niyoruto*	3
	others ∗ ∗ ∗	8
In the previous sentence		13
Subject	company name + *ha*	8
	company name + *ga*	1
	emphasis structure ∗∗	1
	part of the subject ∗	1
Non-subject	company name + *to*	2
In two sentences before		6
Subject	company name + *ha*	5
	company name + *ga*	1
In previous paragraph		1
Topic of the paragraph	company name + *ha*	1
In two paragraphs before		3
Topic of the paragraph	company name + *ha*	3

Table 2 Locations and contexts of the referents of *dousha* with *ha*

dousha with *ha*		
location	context	number of cases
Within the same sentence		2
Subject	company name + *ga*	1
	company name + *deha*	1
In the previous sentence		32
Subject	company name + *ha*	21
	emphasis structure ∗∗	5
	part of the subject ∗	4
Non-subject	others	2
In two sentences before		17
Subject	company name + *ha*	16
	part of the subject ∗	1
In three sentences before (in the same paragraph)		2
Subject	company name + *ha*	2
In previous paragraph		7
Topic of the paragraph	company name + *ha*	6
Topic of the paragraph	company name + *ga*	1
In two paragraphs before		2
Topic of the paragraph	company name + *ha*	2
In three paragraphs before		2
Topic of the paragraph	company name + *ha*	2

Note for Table 1 and Table 2

∗	company name referred to is a part of a larger subject noun phrase.
∗∗	company name referred to comes at the end of the sentence, a way of emphasising the company name in Japanese.
∗ ∗ ∗	company name with *to* (with), *kara* (from), *wo tsuuji* (through), *tono aidade* (between or among).

For *dousha* with *ga* (**Table** 1), the referred company names, or the referents appear in non-subject positions from time to time, especially if the referent appears in the same sentence as *dousha* does. For *dousha* with *ha* (**Table** 2), compared with **Table** 1, very few referents are located in the same sentence, and most of the referents are in the subject position. For both occurrences of *dousha*, a considerable number of the referents appear two or more sentences before, and a few of them show up even two or three paragraphs before.

3 THREE HEURISTIC METHODS TESTED

3.1 Three Heuristic Methods

One simple and two heuristic methods to find the referents of *dousha* are described below. The first, the simple method, is to take the closest company name, (the one which appears most recently before *dousha*), as its referent (**Simple Closest Method** or **SCM**). It is used in this paper to indicate the baseline performance for reference resolution of *dousha*.

The second method is a modified Simple Closest Method for *dousha* with *ga*. It is basically the same as SCM except that:

- if there is one or more company name in the same sentence before the *dousha*, take the closest company name as the referent.
- if there is a company name immediately followed by *ha*, *ga*, *deha*, or *niyoruto* somewhere before *dousha*, use the closest such company name as the referent.
- if the previous sentence ends with a company name, thus putting an emphasis on the company name, make it the referent.
- if there is a pattern "company name *no* human name title..." (equivalent to "title human name of company name..." in English) in the previous sentence, then use the company name as the referent. Typical titles are *shachou* (president) and *kaichou* (Chairman of Board).

The third heuristic method is used for *dousha* with *ha* cases. It is also based on SCM except the following points:

- if there is a company name immediately followed by *ha*, *ga*, *deha*, or *niyoruto* somewhere before *dousha*, use the closest such company name as the referent.
- if the previous sentence ends with a company name, thus putting an emphasis on the company name, make it the referent.
- if there is a pattern "company name *no* human name title..." (equivalent to "title human name of company name..." in English) in the previous sentence, then use the company name as the referent.

The third method is in fact a set of the second method, and both of them use semantic information (i.e. company name, human name, title), syntactic patterns (i.e. where a company name, a human name, or a title appears in a sentence) and several specific lexical items which come immediately after the company names.

3.2 Test Results

The three methods have been tested on the development data from which the methods were produced and on the set of unseen test data.

3.2.1 Against the development data

As mentioned in section two, there are 42 cases of *dousha* with *ga* and 66 cases of *dousha* with *ha*.

For the *dousha* with *ga* cases, the Simple Closest Method identifies the referents **67**% correctly (27 correct out of 42), and the second method does so **90**% (38 out of 42) correctly. SCM misses a number of referents which appear in previous sentences, and most of those which appear two or more sentences previously.

For the cases of *dousha* with *ha*, SCM identifies the referents correctly only **52**% (34 correct out of 66), however, the third heuristic method correctly identifies **94**% (62 out of 66).

3.2.2 Against the test data

The test data was taken from Japanese newspaper articles on micro-electronics. There are 1078 articles, and 51 cases of *dousha* with *ga* and 250 cases of *dousha* with *ha*. The test has been conducted against the all *ga* cases (51 of them) and the first 100 *ha* cases.

For the *dousha* with *ga* cases, the Simple Closest Method identifies the referents **80**% correctly (41 correct out of 51), and the second method does so **96**% (49 out of 51) correctly.

For the cases of *dousha* with *ha*, SCM identifies the referents correctly only **83**% (83 correct out of 100), however, the third heuristic method correctly identifies **96**% (96 out of 100).

The following table, **Table** 3, shows the summary of the test results.

Table 3 Summary of Test Results

	Development Data	Test Data
dousha with *ga*		
SCM	67 %	80 %
2nd method	90 %	96 %
dousha with *ha*		
SCM	52 %	83 %
3rd method	94 %	96 %

4 DISCUSSION

The second and third heuristic methods show high accuracy in finding the referents of *dousha* with *ga* and *ha*. This means that partial semantic parsing (in which key semantic information such as company name, human name, and title is marked) is sufficient for reference resolution of important referential expressions such as *dousha* in Japanese. Moreover, since the two modified methods are simple, they will be easily implemented by computationally inexpensive finite-state pattern matchers (Hobbs *et al.* 1992; Cowie *et al.* 1993). Therefore, they will be suitable for large scale text processing (Jacobs 1992; Chinchor *et al.* 1993).

One important point to realize is that the second and third methods, although they are simple to implement, achieve something that is rather complicated and may be computationally expensive otherwise. For example, in order to find the correct referent of a given *dousha*, you may have to skip one entire paragraph and find the referent two paragraphs before, or you may have to choose the right company name from several possible company names which appear before the given *dousha*. The modified methods do this correctly most of the time without worrying about constructing sometimes complicated syntactic structures of the sentences in the search window for the possible referent.

Another important point is that the modified methods make good use of post-nominal particles, especially *ha* and *ga*. For example, if the referent is located two sentences or more before, then the referent (the company name) comes with *ha* almost all the time (35 out of 38 such cases for both *dousha*). It seems that if the referent of the *dousha* in consideration is more than a certain distance before, two sentences in this case, then the referent is marked with *ha* most of the time. Kitani also uses this *ha* or *ga* marked company names as key information in his reference resolution algorithm for *dousha* (Kitani 1994).

5 CONCLUSION

The locations and contexts of the referents of *dousha* in Japanese Joint-Venture articles are determined by hand. Three heuristic methods are proposed and tested. The methods which use semantic information in the text and its patterns show high accuracy in finding the referents (96% for *dousha* with *ga* and 96% for *dousha* with *ha* for the unseen test data). The high success rates suggest that a semantic pattern-matching approach is not only a valid method but also an efficient method for reference resolution in the newspaper article domains. Since the Japanese language is highly case-inflected, case (particle) information is used effectively in these methods for reference resolution. How much one can do with semantic pattern matching for reference resolution of similar expressions such as "the company" or "the Japanese company" in English newspaper articles is a topic for future research.

6 ACKNOWLEDGEMENT

I would like to thank the Tipster project group at the CRL for their inspiration and suggestions. I would also like to thank Dr. Yorick Wilks, Dr. John Barnden, Mr. Steve Helmreich, and Dr. Jim Cowie for their productive comments. The newspaper articles used in this study are from the Tipster Information Extraction project provided by ARPA.

7 REFERENCES

Chinchor, N., L. Hirschman, and D. Lewis (1993). Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). *Computational Linguistics, 19(3)*, *pp.* 409-449.

Cowie, J., T. Wakao, L. Guthrie, W. Jin, J. Pustejovsky, and S. Waterman (1993). The *Diderot* Information Extraction System. In the proceedings of *The First Conference of the Pacific Association for Computational Linguistics (PACLING 93)* Simon Fraser University, Vancouver, B.C. Canada, *pp.* 23-32.

Jacobs, P.S. (1992). Introduction: Text Power and Intelligent Systems. *In* P.S. Jacobs *Ed.*, *Text-Based Intelligent Systems.* Lawrence Erlbaum Associates, Hillsdale New Jersey, *pp.* 1-8.

Hobbs, J., D. Appelt, M. Tyson, J. Bear, and D. Israel (1992). SRI International Description of the FASTUS System used for MUC-4. In the proceedings of *Fourth Message Understanding Conference (MUC-4)*, Morgan Kaufmann Publishers, San Mateo, *pp.* 269-275.

Kitani, T. (1994). Merging Information by Discourse Processing for Information Extraction. In the proceedings of *the tenth IEEE Conference on Artificial Intelligence for Applications*, *pp.* 168-173.

Muraki, K., S. Doi, and S. Ando (1993). Context Analysis in Information Extraction System based on Keywords and Text Structure. In the proceedings of *the 47th National Conference of Information Processing Society of Japan*, *3-81.* (In Japanese).

Shibata, M., O. Tanaka, and J. Fukumoto (1990). Anaphora in Newspaper Editorials. In the proceedings of *the 40th National Conference of Information Processing Society of Japan*, *5F-4.* (In Japanese).